# Asymptotic connectivity for the network of RNA secondary structures

P. Clote

Biology Department, Boston College, Chestnut Hill, MA 02467, clote@bc.edu

**Abstract**

Given an RNA sequence $\mathbf{a}$, consider the network $G = (V, E)$, where the set $V$ of nodes consists of all secondary structures of $\mathbf{a}$, and whose edge set $E$ consists of all edges connecting two secondary structures whose base pair distance is 1. Define the network *connectivity*, or *expected network degree*, as the average number of edges incident to vertices of $G$. Using algebraic combinatorial methods, we prove that the asymptotic connectivity of length $n$ homopolymer sequences is $0.473418 \cdot n$. This raises the question of what other network properties are characteristic of the network of RNA secondary structures. Programs in Python, C and Mathematica are available at the web site http://bioinformatics.bc.edu/clotelab/RNAexpNumNbors.

## 1 Introduction

In [24], Stein and Waterman used generating function theory to determine the asymptotic number of RNA secondary structures. Since that pioneering paper, a number of results concerning RNA secondary structure asymptotics have appeared, including the incomplete list [13, 12, 19, 3, 17, 5, 11, 15, 2, 22, 6, 23, 10, 16, 8].

In contrast to the previous papers, here we consider *network* properties of the ensemble of RNA secondary structures, following the seminal work of Wuchty [30], who showed by exhaustive enumeration of the low energy secondary structures of *E. coli* phe-tRNA, that the corresponding network architecture had *small-world* properties [29]. Small-world networks appear to abound in biology, providing a kind of robustness necessary for the molecular processes of life, as seen in networks of neural connections of *C. elegans* [29], gene co-expression in *S. cerevisiae* [25], metabolic pathways [26, 21], intermediate conformations in tertiary folding kinetics for the protein villin [1], etc.

In this paper, we use algebraic combinatorial methods, and in particular the Flajolet-Odlyzko theorem [7] to prove that the asymptotic expected network connectivity of RNA secondary structures is $0.47341764431521986 \cdot n$. Following [28, 14], *stickiness* is defined to be the probability $p$ that any two positions can pair. For the simplicity of argument, in the homopolymer model, we take stickiness $p$ to be 1; however, minor changes in our C dynamic programming algorithm and in our Mathematica code permit the computation of asymptotic expected connectivity for arbitrary stickiness $p$.

## 2 Preliminaries

A secondary structure for a given RNA nucleotide sequence $\mathbf{a} = \mathbf{a}_1, \ldots, \mathbf{a}_n$ is a set $s$ of base pairs $(i, j)$, such that *(i)* if $(i, j) \in s$ then $\mathbf{a}, \mathbf{a}_j$ form either a Watson-Crick (AU,UA,CG,GC) or wobble (GU,UG) base pair, *(ii)* if $(i, j) \in s$ then $j - i > \theta = 3$ (a steric constraint requiring that there be at least $\theta = 3$ unpaired bases between any two paired bases), *(iii)* if $(i, j) \in s$ then for all $j' \neq j$ and $i' \neq i$, $(i', j) \notin s$ and $(i, j') \notin s$ (nonexistence of base triples), *(iv)* if $(i, j) \in s$ and $(k, \ell) \in s$, then it is not the case that $i < k < j < \ell$ (nonexistence of pseudoknots). For the purposes of this paper, following Stein and Waterman [24], we consider the *homopolymer* model of RNA, in which condition *(i)* is dropped, so that any base can pair with any other base.

Suppose that $\mathbf{a} = \mathbf{a}_1, \ldots, \mathbf{a}_n$ is an RNA sequence. If $s$ is a secondary structure of $\mathbf{a}$, then let $N_s$ denote the number of secondary structures of $\mathbf{a}$ that can be obtained from $s$ by the removal or addition of a single

base pair; i.e. those structures having base pair distance from $s$ of 1. Define the *expected number of neighbors* $\langle N_s \rangle$ by

$$\langle N_s \rangle \quad = \quad \frac{Q}{Z} \tag{1}$$

where $Q = \sum_s N_s$ is the total number $N_x$ of neighbors of all secondary structures $s$ of $\mathbf{a}$, and $Z$ denotes the total number of secondary structures of $\mathbf{a}$. Note that $Z$ corresponds to the *partition function* $\sum_s \exp(-E(s)/RT)$ if the energy of every structure is 0.

In [4] we described three algorithms to compute the expected number of neighbors, or *network connectivity*, $\langle N_s \rangle = \sum_s \frac{\exp(-E(s)/RT)}{Z} \cdot N_s$, where $Z$ is the partition function $\sum_s \exp(-E(s)/RT)$ with respect to energy model A (each structure $s$ has energy $E(s) = 0$), model B (each structure $s$ has Nussinov [20] energy $E(s)$ equal to $-1$ times the number $|s|$ of base pairs in $s$), and model C (each structure $s$ has energy $E(s)$ given by the Turner energy model [18, 31]).

Below, we follow reference [4] in deriving the recurrence relations for $Q$ and $Z$ for model A, corresponding to equation equation (1). For $1 \leq i \leq j \leq n$, define the subsequence $\mathbf{a}[i,j] = \mathbf{a}_i, \ldots, \mathbf{a}_j$, and define $SS(\mathbf{a}[i,j])$ to be the collection of secondary structures of $\mathbf{a}[i,j]$. Define

$$Q_{i,j} = \sum_{s \in SS(\mathbf{a}[i,j])} N_s. \tag{2}$$

Similarly, let $Z_{i,j} = \sum_{s \in SS(\mathbf{a}[i,j])} 1$; i.e. $Z_{i,j}$ denotes the number of secondary structures of $\mathbf{a}[i,j]$.

Base Case: For $j - i \in \{0, 1, 2, 3\}$, $Q_{i,j} = 0$ and $Z_{i,j} = 1$.

Inductive Case: Let $BP(i, j, \mathbf{a})$ be a boolean function, taking the value 1 if positions $i, j$ can form a base pair for sequence $\mathbf{a}$, and otherwise taking the value 0. Assume that $j - i > 3$.

Subcase A: Consider all secondary structures $s \in \mathbf{a}[i,j]$, for which $j$ is unpaired. For each structure $s$ in this subcase, the number $N_s$ of neighbors of $s$ is constituted from the number of structures obtained from $s$ by removal of a single base pair, together with the number of structures obtained from $s$ by addition of a single base pair. If the base pair added does not involve terminal position $j$, then total contribution to $\sum_{s \in SS(\mathbf{a}[i,j])} N_s$ is $Q_{i,j-1}$. It remains to count the contribution due to neighbors $t$ of $s$, obtained from $s \in SS(\mathbf{a}[i,j])$ by adding the base pair $(k, j)$. This contribution is given by $\sum_{k=i}^{j-4} BP(k, j, \mathbf{a}) \cdot Z_{i,k-1} \cdot Z_{k+1,j-1}$, where $Z_{i,i-1}$ is defined to be 1. Thus the total contribution to $Q_{i,j}$ from this subcase is

$$Q_{i,j-1} + \sum_{k=i}^{j-4} BP(k, j, \mathbf{a}) \cdot Z_{i,k-1} \cdot Z_{k+1,j-1}.$$

Subcase B: Consider all secondary structures $s \in \mathbf{a}[i,j]$ that contain the base pair $(k, j)$ for some $k \in \{i, \ldots, j-4\}$. For secondary structure $s$ in this subcase, the number $N_s$ of neighbors of $s$ is constituted from the number of structures obtained by removing base pair $(k, j)$ together with a contribution obtained by adding/removing a single base pair either to the region $[i, k-1]$ or to the region $[k+1, j-1]$. Setting $Q_{i,i-1}$ to be 0, these contributions are given by

$$\sum_{k=i}^{j-4} BP(k, j, \mathbf{a}) \cdot \left[ Z_{i,k-1} \cdot Z_{k+1,j-1} + Q_{i,k-1} \cdot Z_{k+1,j-1} + Z_{i,k-1} \cdot Q_{k+1,j-1} \right].$$

In the current subcase, the contribution to $Z_{i,j}$ is $\sum_{k=i}^{j-4} BP(k, j, \mathbf{a}) \cdot Z_{i,k-1} \cdot Z_{k+1,j-1}$.

Finally, taking the contributions from both subcases together, it follows that

$$Q_{i,j} \quad = \quad Q_{i,j-1} + \sum_{k=i}^{j-4} BP(k, j, \mathbf{a}) \cdot [2 \cdot Z_{i,k-1} \cdot Z_{k+1,j-1} + Q_{i,k-1} \cdot Z_{k+1,j-1} + Z_{i,k-1} \cdot Q_{k+1,j-1}] \tag{3}$$

$$Z_{i,j} \quad = \quad Z_{i,j-1} + \sum_{k=i}^{j-4} BP(k, j, \mathbf{a}) \cdot Z_{i,k-1} \cdot Z_{k+1,j-1}. \tag{4}$$

```
a  .......   6
b  (...)..   1
c  (....).   1
d  .(...).   2
e  (.....)   2
f  .(....)   1
g  ..(...)   1
h  ((...))   2
```

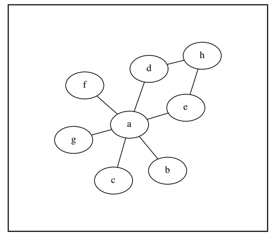


Figure 1: *(Left)* All possible secondary structures of the 7-mer homopolymer, where position $i$ can pair with position $j$ provided only that $1 \leq i < j \leq 7$ and $j - i \geq 4$. *(Right)* Graph representation of neighborhood network, where nodes a,b,c,d,e,f,g,h respectively represent the 8 secondary structures in the list. The number of neighbors of each secondary structure is indicated to its right. The expected number of neighbors for the 7-mer is thus $(6+1+1+2+2+1+1+2)/8 = 16/8 = 2$.

It follows that the expected number $\langle N_s \rangle$ of neighbors $N_s$ of structures $s$ of $\mathbf{a}$ is $\frac{Q_{1,n}}{Z_{1,n}}$. Note that the recursion for $Z_{i,j}$ is well-known and due originally to Waterman [27].

To provide concrete intuition for the problem we consider, in Figure 1, we present the list of secondary structures for a homopolymer of length 7, depicted as a network having expected connectivity of 2. In the left panel of Figure 2, we present a histogram for the network connectivity (graph degree or number of neighbors), by analyzing an exhaustively produced list of all 106,633 structures of the 20-mer homopolymer. In the right panel of Figure 2, we present a plot of the *normalized* expected number of neighbors of for homopolymers of length 1 to 1000 nt, obtained by dividing the expected number of neighbors by sequence length. Clearly there appears to be an asymptotic value for the length-normalized expected connectivity, suggesting that it may be possible to formally prove the existence of this asymptotic value, a task to which the remainder of the paper is dedicated.

# 3 Methods

## 3.1 Recurrence relation for partition function $Z_n$

In order to determine asymptotic network connectivity, we now provide similar recursions to those of (3) and (4) for the *homopolymer* model of RNA, where any base (position) $i$ can pair with any other position (base) $j$, provide only that $1 \leq i + \theta + 1 \leq j \leq n$. Following common convention due to steric constraints, we take $\theta$ to be 3. These recursions are the basis of the dynamic programming code we implemented, used to produce Figure 2.

We define $Z_0 = 1$, in order to simplify the recurrence relation for $Z_n$, defined to be the number of secondary structures for a homopolymer of length $n$, or equivalently, the *partition function* for the energy model that assigns an energy of 0 to every structure. Moreover, since the empty structure is the only structure for sequences of length $1, 2, 3, 4 = \theta + 1$, we define $Z_n = 1$ for $0 \leq n \leq 4$. Secondary structures for a homopolymer of length $n > 4$ can be partitioned into two classes: (1) $n$ is unpaired, (2) there is a base pair $(k, n)$ for some $1 \leq k \leq n - 4 = n - \theta - 1$. Thus we have

$$Z_n = \begin{cases} 1 & \text{if } 0 \leq n \leq 4 = \theta + 1 \\ Z_{n-1} + \sum_{k=0}^{n-\theta-2} Z_k \cdot Z_{n-2-k} & \text{if } n \geq 5 = \theta + 2 \end{cases} \tag{5}$$

which is the homopolymer analogue of equation (4). To employ generating function theory, we require a single formula for $Z_n$, rather than a definition by cases – see p. 66 of [9]. This is easily achieved by *(i)* adding the indicator function $I[n = 0]$, defined to equal 1 if $n = 0$, and otherwise 0, and *(ii)* adding and subtracting the same terms to ensure that $k$ ranges from 0 to $n - 2$, rather than $n - \theta - 2 = n - 5$. Thus

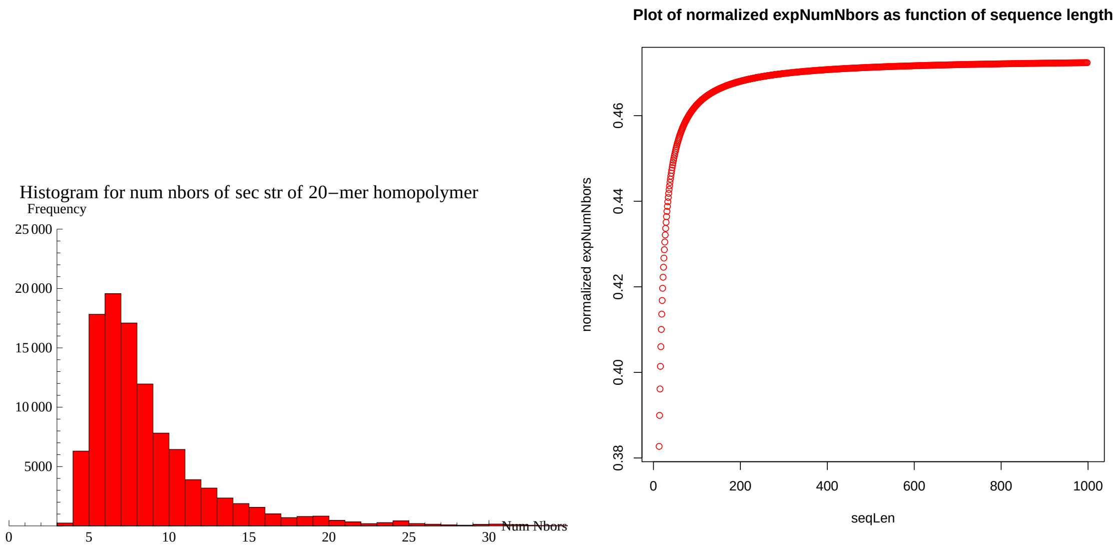


Figure 2: *(Left)* Histogram of the number of neighbors for all 106,633 secondary structures of the 20-mer homopolymer. The mean is 8.336, the standard deviation is 4.769, the maximum is 136, and minimum is 3. *(Right)* Plot of the *normalized* expected number of neighbors of for homopolymers of length 1 to 1000 nt, obtained by dividing the expected number of neighbors by sequence length. Apparent asymptotic value seems to be $\approx 0.4724$. The main result of this paper is the proof that the asymptotic value is in fact 0.4734176431521986.

equation (5) is equivalent to (6) defined by:

$$\begin{aligned} Z_n &= Z_{n-1} + \sum_{k=0}^{n-2} Z_k \cdot Z_{n-2-k} + I[n=0] - Z_{n-1} \cdot Z_0 - Z_{n-3} \cdot Z_1 - Z_{n-4} \cdot Z_2 \\ &= Z_{n-1} + \sum_{k=0}^{n-2} Z_k \cdot Z_{n-2-k} + I[n=0] - Z_{n-1} - Z_{n-3} - Z_{n-4}. \end{aligned} \tag{6}$$

Let $z = \sum_{n=0}^{\infty} Z_n \cdot x^n$. By multiplying equation (6) by $x^n$, summing from $n = 0$ to $\infty$, we obtain

$$\sum_{n=0}^{\infty} Z_n \cdot x^n = \sum_{n=0}^{\infty} Z_{n-1} \cdot x^n + \sum_{n=0}^{\infty} \sum_{k=0}^{n-2} Z_k \cdot Z_{n-2-k} \cdot x^n - \sum_{n=0}^{\infty} (Z_{n-2} + Z_{n-3} + Z_{n-4} - I[n=0]) \cdot x^n$$

which yields

$$z = xz + x^2 z^2 - zx^2 - zx^3 - zx^4 + 1 \tag{7}$$

Solving the quadratic equation (7) for $z$, we determine that

$$z = \frac{1 - x + x^2 + x^3 + x^4 \pm \sqrt{1 - 2x - x^2 + x^4 + 3x^6 + 2x^7 + x^8}}{2x^2}.$$

Only the first solution

$$z1 = \frac{1 - x + x^2 + x^3 + x^4 - \sqrt{1 - 2x - x^2 + x^4 + 3x^6 + 2x^7 + x^8}}{2x^2} \tag{8}$$

has the property that the coefficients of its Taylor expansion correspond to the values of $Z_n$, as determined by dynamic programming (see program at web site). In particular, using Mathematica, we obtain

$$\begin{aligned} z1 = \; & 1 + x + x^2 + x^3 + x^4 + 2x^5 + 4x^6 + 8x^7 + 16x^8 + 32x^9 + 65x^{10} + 133x^{11} + 274x^{12} + 568x^{13} + \\ & 1184x^{14} + 2481x^{15} + 5223x^{16} + 11042x^{17} + 23434x^{18} + 49908x^{19} + 106633x^{20} + O(x)^{21} \end{aligned}$$

which can be compared with the values determined by our dynamic programming implementation:

| $n$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Z_n$ | 1 | 1 | 1 | 1 | 1 | 2 | 4 | 8 | 16 | 32 | 65 | 133 | 274 | 568 | 1184 | 2481 | 5223 | 11042 |

### 3.2 Recurrence relation for the number of neighbors $Q_n$

Let $Q_n$ denote the total number $N_s$ of nearest neighbors of all secondary structures $s$ of a homopolymer of length $n$. The homopolymer analogue of equation (3) is as follows:

$$Q_n = \begin{cases} 0 & \text{if } n = 0,1,2,3,4 \\ Q_{n-1} + \sum_{k=0}^{n-\theta-2} 2Z_k Z_{n-2-k} + Q_k Z_{n-2-k} + Z_k Q_{n-2-k} & \text{else.} \end{cases} \tag{9}$$

If we assume that $Z_n = 0 = Q_n$ for $n < 0$, then it follows from equations (9) and (5), that we can express $Q_n$ by the formula

$$Q_n = Q_{n-1} + \sum_{k=0}^{n-5} 2Z_k Z_{n-2-k} + Q_k Z_{n-2-k} + Z_k Q_{n-2-k}. \tag{10}$$

Next, we add and subtract the same terms in order to ensure that the upper bound in the previous summation is $n-2$. This yields

$$\begin{aligned} Q_n = {} & Q_{n-1} + \sum_{k=0}^{n-2} 2Z_k Z_{n-2-k} + Q_k Z_{n-2-k} + Z_k Q_{n-2-k} \\ & -2\left(Z_{n-2}\cdot Z_0 + Z_{n-3}\cdot Z_1 + Z_{n-4}\cdot Z_2\right) - \left(Q_{n-2}\cdot Z_0 + Q_{n-3}\cdot Z_1 + Q_{n-4}\cdot Z_2\right) \\ & -\left(Z_{n-2}\cdot Q_0 + Z_{n-3}\cdot Q_1 + Z_{n-4}\cdot Q_2\right) \end{aligned} \tag{11}$$

which simplifies to

$$\begin{aligned} Q_n = {} & Q_{n-1} + \sum_{k=0}^{n-2} 2Z_k Z_{n-2-k} + Q_k Z_{n-2-k} + Z_k Q_{n-2-k} \\ & -2\left(Z_{n-2} + Z_{n-3} + Z_{n-4}\right) - \left(Q_{n-2} + Q_{n-3} + Q_{n-4}\right). \end{aligned} \tag{12}$$

Multiply each term of equation (12) by $x^n$ and summing from $n = 0$ to $\infty$, to obtain

$$\begin{aligned} & \sum_{n=0}^{\infty} Q_n \cdot x^n = \sum_{n=0}^{\infty} Q_{n-1}\cdot x^n + \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n-2} 2Z_k Z_{n-2-k}\right) + \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n-2} Q_k Z_{n-2-k}\right) + \\ & \sum_{n=0}^{\infty}\left(\sum_{k=0}^{n-2} Z_k Q_{n-2-k}\right) - \sum_{n=0}^{\infty} 2\left(Z_{n-2} + Z_{n-3} + Z_{n-4}\right) - \sum_{n=0}^{\infty}\left(Q_{n-2} + Q_{n-3} + Q_{n-4}\right). \end{aligned} \tag{13}$$

Let $q = \sum_{n=0}^{\infty} Q_n x^n$ and $z = \sum_{n=0}^{\infty} Z_n x^n$. Then from (13) we have

$$q = xq + 2x^2z^2 + x^2qz + x^2zq - 2x^2z - 2x^3z - 2x^4z - qx^2 - qx^3 - qx^4. \tag{14}$$

From equations (7) and (14), we have

- $z^2x^2 = z - zx + zx^2 + zx^3 + zx^4 - 1$
- $q = xq + 2x^2z^2 + 2x^2qz - 2x^2z - 2x^3z - 2x^4z - qx^2 - qx^3 - qx^4$

from which we eliminate the variable $z$ to obtain the following quadratic equation in variable $q$,

$$\begin{aligned} 4x^5 = {} & q^2x^2\left(1 - 2x - x^2 + x^4 + 3x^6 + 2x^7 + x^8\right) + \\ & q\left(2 - 6x + 2x^2 + 2x^3 + 2x^4 - 2x^5 + 6x^6 - 2x^7 - 2x^8 - 2x^9\right). \end{aligned} \tag{15}$$

Solving for $q$, we determine that only the solution

$$\begin{aligned} q2 &= \left(-1+3x-x^2-x^3-x^4+x^5-3x^6+x^7+x^8+x^9+\right. \\ &\sqrt{\left(1-6x+11x^2-4x^3-3x^4-6x^5+15x^6-16x^7+x^8+2x^9+9x^{10}-\right.} \\ &\left.\left.8x^{11}+7x^{12}+6x^{13}+5x^{14}+3x^{16}+2x^{17}+x^{18}\right)\right) / \left(x^2-2x^3-x^4+x^6+3x^8+2x^9+x^{10}\right) \end{aligned} \tag{16}$$

is possible, since its Taylor expansion is

$$\begin{aligned} q2 &= 2x^5+6x^6+16x^7+40x^8+96x^9+228x^{10}+532x^{11}+1230x^{12}+2826x^{13}+6464x^{14}+14742x^{15}+ \\ &33546x^{16}+76216x^{17}+172968x^{18}+392228x^{19}+888932x^{20}+O(x)^{21}. \end{aligned}$$

These values agree with those from the following table that were obtained by our dynamic programming implementation.

| $n$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Q_n$ | 0 | 0 | 0 | 0 | 0 | 2 | 6 | 16 | 40 | 96 | 228 | 532 | 1230 | 2826 | 6464 | 14742 | 33546 | 76216 |

## 3.3 Flajolet-Odlyzko theorem

Having determined the formulas for $z1$ [resp. $q2$] in equation (8) [resp. equation (16), we use the following theorem of Flajolet and Odlyzko [7] to determine the asymptotic coefficients of $Z_n$ [resp. $Q_n$] in the Taylor expansion of the generating function $z1$ [resp. $q2$].

Following standard convention, let $[x^n]f(x)$ denote the $n$th coefficient in the Taylor expansion of $f(x)$. The following theorem is stated as Corollary 2, part (i) of [7] on page 224.

**Theorem 1 (Flajolet and Odlyzko)** *Assume that $f(x)$ has a singularity at $x = \rho > 0$, is analytic in the rest of the region $\triangle\backslash 1$, depicted in Figure 3, and that as $x \to \rho$ in $\triangle$,*

$$f(x) \sim K(1 - x/\rho)^{\alpha}. \tag{17}$$

*Then, as $n \to \infty$, if $\alpha \notin 0, 1, 2, \ldots$,*

$$f_n = [x^n]f(x) \sim \frac{K}{\Gamma(-\alpha)} \cdot n^{-\alpha-1}\rho^{-n}$$

*where $\Gamma$ denotes the Gamma function.*

In Section 3.4, we determine the asymptotic value of $[x^n]q2 = Q_n$, and in the following Section 3.5, we determine the asymptotic value of $[x^n]z1 = Z_n$. The ratio of these values then yields the asymptotic expected number of neighbors, or *network connectivity* for the homopolymer model of RNA.

## 3.4 Asymptotic number of neighbors

Let $P$ denote the polynomial under the radical of equation (16), i.e.

$$P = 1-6x+11x^2-4x^3-3x^4-6x^5+15x^6-16x^7+x^8+2x^9+ \tag{18}$$
$$9x^{10}-8x^{11}+7x^{12}+6x^{13}+5x^{14}+3x^{16}+2x^{17}+x^{18}. \tag{19}$$

There are 4 real roots and 14 imaginary roots of $P$; however, the root of the smallest modulus (absolute value) is the real root

$$\rho = 0.436911127214519 \approx 0.436911. \tag{20}$$

Now $q2$ can be expressed in the form $q2 = G + H$, where $G = \frac{Gnum}{Gdenom}$ and $H = \frac{\sqrt{P}}{Gdenom}$, and where

$$\begin{aligned} Gnum &= \left(-1+3x-x^2-x^3-x^4+x^5-3x^6+x^7+x^8+x^9\right) \\ Gdenom &= x^2\left(1-2x-x^2+x^4+3x^6+2x^7+x^8\right). \end{aligned} \tag{21}$$

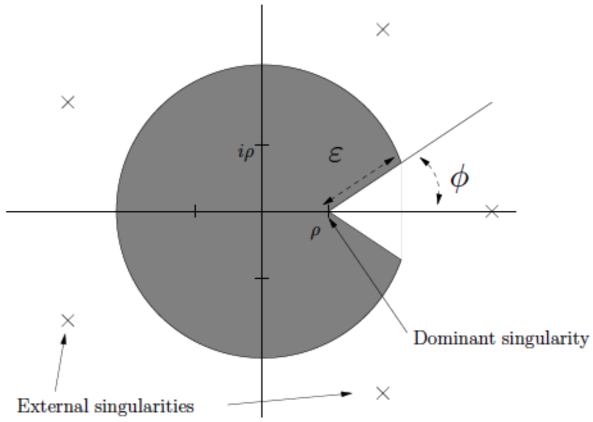


Figure 3: The shaded region $\triangle$ where, except at $x = \rho$, the generating function $f(x)$ must be analytic. Here $\rho = 1$. Figure taken from Lorenz et al. [17].

One notes that $\rho$ is a root of both $Gnum$ and $Gdenom$, so their ratio is well-defined; as well, clearly 0 is a root of $Gdenom$. It follows that both 0 and $\rho$ are singularities of the function $q2$ (recall that in complex analysis, the square root of 0 is a singularity).

For asymptotics, the term $\frac{Gnum}{Gdenom}$ can be neglected, since $\lim_{x\to\rho} \frac{Gnum}{Gdenom} = -2.963617606602476$; thus it follows that $[x^n]q2 = [x^n]\frac{\sqrt{P}}{Gdenom}$. However, the Flajolet-Odlyzko theorem cannot be applied to the function $f = \frac{\sqrt{P}}{Gdenom}$, since $\rho$ is not the dominant singularity, as $0 < \rho$. To address this issue, we define $Gnum1 = Gnum$ and $Gdenom1 = Gdenom/x^2$, hence

$$\begin{aligned} Gnum1 &= \left(-1 + 3x - x^2 - x^3 - x^4 + x^5 - 3x^6 + x^7 + x^8 + x^9\right) \\ Gdenom1 &= \left(1 - 2x - x^2 + x^4 + 3x^6 + 2x^7 + x^8\right). \end{aligned} \tag{22}$$

Now we can apply Theorem 1 to the function $f1 = x^2 \cdot f = \frac{\sqrt{P}}{Gdenom1}$, for which $\rho$ is the dominant singularity. First, we factor $(1 - x/\rho)$ out from $P$.

$$\frac{P}{1 - x/\rho} = 1 - 3.71121x + 2.50581x^2 + 1.73529x^3 + 0.971726x^4 - 3.77592x^5 + 6.3577x^6 - \tag{23}$$

$$1.44854x^7 - 2.3154x^8 - 3.29948x^9 + 1.44816x^{10} - 4.68545x^{11} - 3.72404x^{12} - \tag{24}$$

$$2.52356x^{13} - 0.775919x^{14} - 1.77592x^{15} - 1.06471x^{16} - 0.436911x^{17}. \tag{25}$$

Second, we factor $(1 - x/\rho)$ out from $Gdenom1$.

$$\begin{aligned} \frac{Gdenom1}{1 - x/\rho} &= 1 + 0.288795x - 0.339007x^2 - 0.775919x^3 - 0.775919x^4 - 1.77592x^5 - \\ &\quad 1.06471x^6 - 0.436911x^7. \end{aligned} \tag{26}$$

It follows from (23) and (26) that

$$\frac{\sqrt{P}}{Gdenom1} = \frac{(1 - x/\rho)^{-1/2} \cdot \sqrt{P/(1 - x/\rho)}}{Gdenom1/(1 - x/\rho)} \tag{27}$$

$$\frac{\sqrt{P/(1 - x/\rho)}}{Gdenom1/(1 - x/\rho)}(\rho) = 0.11422693623949792. \tag{28}$$

and so $\frac{\sqrt{P}}{Gdenom1} = 0.11422693623949792 \cdot (1 - x/\rho)^{-1/2}$. If we define $K = 0.11422693623949792$, then it follows that $f1 = \frac{\sqrt{P}}{Gdenom1} \sim K(1 - x/\rho)^{-1/2}$, and by applying Theorem 1 for $\alpha = -1/2$, we have the following asymptotic result.

**Lemma 1** *The asymptotic value of the nth coefficient* $[x^n]\left(x^2 q2\right)$ *in the Taylor expansion of* $x^2 \cdot \sum_{n=0}^{\infty} Q_n x^n$ *is* $\frac{0.06444564758689844 \cdot 2.2887949921884863^n}{\sqrt{n}}$.

PROOF. We have

$$
\begin{aligned}
[x^n]\left(x^2 q2\right) &= [x^n]\,(f1) \sim \frac{K}{\Gamma(-\alpha)} \cdot n^{-\alpha-1} \cdot \rho^{-n} \\
&= \frac{0.11422693623949792}{\Gamma(1/2)} \cdot n^{-1/2} \cdot 0.436911127214519^{-n} \\
&= \frac{0.06444564758689844 \cdot 2.2887949921884863^n}{\sqrt{n}} \square
\end{aligned} \tag{29}
$$

## 3.5 Asymptotic number of structures

We now proceed similarly to determine the asymptotic value of the Taylor coefficients of $x^2 \cdot \sum_{n=0}^{\infty} Z_n x^n$. Now let $P$ denote the polynomial under the radical of equation (8), i.e.

$$
P = 1 - 2x - x^2 + x^4 + 3x^6 + 2x^7 + x^8. \tag{30}
$$

There are 2 real roots and 6 imaginary roots of $P$; however, the root of the smallest modulus (absolute value) is the real root

$$
\rho = 0.436911127214519 \approx 0.436911, \tag{31}
$$

identical the the value from equation (20). Then $z1$ can be expressed in the form $z1 = G + H$, where $G = \frac{Gnum}{Gdenom}$ and $H = \frac{\sqrt{P}}{Gdenom}$, and where

$$
\begin{aligned}
Gnum &= \left(1 - x + x^2 + x^3 + x^4\right) \\
Gdenom &= 2x^2.
\end{aligned} \tag{32}
$$

Note that $\rho$ is not a root of either $Gnum$ or $Gdenom$, so their ratio is well-defined; as well, clearly 0 is a root of $Gdenom$. It follows that both 0 and $\rho$ are singularities of the function $z1$ (recall that in complex analysis, the square root of 0 is a singularity).

For asymptotics, the term $\frac{Gnum}{Gdenom}$ can be neglected, since $\lim_{x\to\rho} \frac{Gnum}{Gdenom} = 2.2887949921884205$; thus it follows that $[x^n]z1 = [x^n]\frac{\sqrt{P}}{Gdenom}$. However, the Flajolet-Odlyzko theorem cannot be applied to the function $f = \frac{\sqrt{P}}{Gdenom}$, since 0, rather than $\rho$, is the dominant singularity. To address this issue, we define $Gnum1 = Gnum$ and $Gdenom1 = Gdenom/x^2$, hence

$$
\begin{aligned}
Gnum1 &= \left(1 - x + x^2 + x^3 + x^4\right) \\
Gdenom1 &= 2.
\end{aligned} \tag{33}
$$

Now we can apply Theorem 1 to the function $f1 = x^2 \cdot f = \frac{\sqrt{P}}{Gdenom1}$, for which $\rho$ is the dominant singularity. First, we factor $(1 - x/\rho)$ out from $P$.

$$
\frac{P}{1 - x/\rho} = 1 + 0.288795x - 0.339007x^2 - 0.775919x^3 - 0.775919x^4 - 1.77592x^5 - 1.06471x^6 - 0.436911x^7 \tag{34}
$$

It follows from (34) that

$$
\frac{\sqrt{P}}{Gdenom1} = \frac{(1 - x/\rho)^{1/2} \cdot \sqrt{P/(1 - x/\rho)}}{Gdenom1} \tag{35}
$$

$$
\frac{\sqrt{P/(1 - x/\rho)}}{Gdenom1}(\rho) = \frac{\sqrt{P/(1 - x/\rho)}}{2}(\rho) = 0.4825630725501931 \tag{36}
$$

and so $\frac{\sqrt{P}}{2} = 0.4825630725501931 \cdot (1 - x/\rho)^{1/2}$. If we define $K = 0.4825630725501931$, then it follows that $f1 = \frac{\sqrt{P}}{2} \sim K(1 - x/\rho)^{1/2}$, and by applying Theorem 1 for $\alpha = +1/2$, we have the following asymptotic result.

**Lemma 2** *The asymptotic value of the nth coefficient* $[x^n]\left(x^2 z1\right)$ *in the Taylor expansion of* $x^2 \cdot \sum_{n=0}^{\infty} Z_n x^n$ *is* $\frac{0.13612852946880957 \cdot 2.2887949921884863^n}{n^{3/2}}$.

PROOF. We have

$$\begin{aligned}
[x^n]\left(x^2 z1\right) &= [x^n]\,(f1) \sim \frac{K}{\Gamma(-\alpha)} \cdot n^{-\alpha-1} \cdot \rho^{-n} \\
&= \frac{0.13612852946880957}{\Gamma(-1/2)} \cdot n^{-3/2} \cdot 2.2887949921884863^n \\
&= \frac{0.13612852946880957 \cdot 2.2887949921884863^n}{n^{3/2}}. \qquad (37)
\end{aligned}$$

□ We have now established the asymptotic *expected* network connectivity.

**Theorem 2** *The asymptotic value of the expected number of neighbors is* $0.4734176431521986 \cdot n$*; i.e. the asymptotic value of* $\sum_{n=0}^{\infty} \frac{Q_n}{nZ_n} \cdot x^n$ *is* $0.4734176431521986$.

PROOF. By Lemmas 1 and 2, we have

$$\begin{aligned}
\frac{[x^n]q2}{[x^n]z1} &= \frac{[x^n](x^2 \cdot q2)}{[x^n](x^2 \cdot z1)} \\
&= \frac{(0.06444564758689844 \cdot 2.2887949921884863^n \cdot n^{-1/2}}{0.13612852946880957 \cdot 2.2887949921884863^n \cdot n^{-3/2}} \\
&= 0.4734176431521986 \cdot n.
\end{aligned}$$

# 4 Acknowledgements

This research was funded by the National Science Foundation grant DBI-1262439. Akademischer Austauschdienst. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

# References

[1] G. R. Bowman and V. S. Pande. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U.S.A.*, 107(24):10890–10895, June 2010.

[2] W.Y. C. Chen, J. Qin, C. Reidys, and D. Zeilberger. Efficient counting and asymptotics of k-noncrossing tangled diagrams. *Electr. J. Comb.*, 16(1), 2009.

[3] P. Clote. Combinatorics of saturated secondary structures of RNA. *J. Comput. Biol.*, 13(9):1640–1657, November 2006.

[4] P Clote. Expected degree for RNA secondary structure networks. *J Comp Chem*, 2014.

[5] P. Clote, E. Kranakis, D. Krizanc, and B. Salvy. Asymptotics of canonical and saturated RNA secondary structures. *J. Bioinform. Comput. Biol.*, 7(5):869–893, October 2009.

[6] P. Clote, Y. Ponty, and J. M. Steyaert. Expected distance between terminal nucleotides of RNA secondary structures. *J Math Biol.*, 0(O):O, October 2011.

[7] P. Flajolet and A. M. Odlyzko. Singularity analysis of generating functions. *SIAM Journal of Discrete Mathematics*, 3:216–240, 1990.

[8] E. Fusy and P. Clote. Combinatorics of locally optimal RNA secondary structures. *J Math Biol.*, 0(O):O, December 2012.

[9] R.L Graham, D.E. Knuth, and O. Patashnik. *Concrete Mathematics - A Foundation for Computer Science*. Addison-Wesley, 1989.

[10] H. S. Han and C. M. Reidys. The 5'-3' distance of RNA secondary structures. *J. Comput. Biol.*, 19(7):867–878, July 2012.

[11] Hillary S. W. Han and Christian M. Reidys. Stacks in canonical RNA pseudoknot structures. *Mathematical Biosciences*, 219(1):7–14, May 2009.

[12] Christian Haslinger and Peter F. Stadler. Rna structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bulletin of Mathematical Biology*, 61(3):437–467, May 1999.

[13] Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discrete Appl. Math.*, 88(1-3):207–237, 1998.

[14] Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 88:207–237, 1998.

[15] Emma Y. Jin and Christian M. Reidys. On the decomposition of k-noncrossing RNA structures. *Advances in Applied Mathematics*, 44(1):53–70, January 2010.

[16] T. J. Li and C. M. Reidys. Combinatorics of RNA-RNA interaction. *J. Math. Biol.*, 64(3):529–556, February 2012.

[17] W. A. Lorenz, Y. Ponty, and P. Clote. Asymptotics of RNA shapes. *J. Comput. Biol.*, 15(1):31–63, 2008.

[18] D.H. Matthews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[19] M. E. Nebel. Combinatorial properties of RNA secondary structures. *J. Comput. Biol.*, 9(3):541–573, 2002.

[20] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences, USA*, 77(11):6309–6313, 1980.

[21] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, August 2002.

[22] C. M. Reidys, F. W. Huang, J. E. Andersen, R. C. Penner, P. F. Stadler, and M. E. Nebel. Topology and prediction of RNA pseudoknots. *Bioinformatics*, 27(8):1076–1085, April 2011.

[23] C. Saule, M. Regnier, J. M. Steyaert, and A. Denise. Counting RNA pseudoknotted structures. *J. Comput. Biol.*, 18(10):1339–1351, October 2011.

[24] P. R. Stein and M. S. Waterman. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Mathematics*, 26:261–272, 1978.

[25] V. Van Noort, B. Snel, and M. A. Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.*, 5(3):280–284, March 2004.

[26] A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proc. Biol. Sci.*, 268(1478):1803–1810, September 2001.

[27] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Studies in Foundations and Combinatorics, Advances in Mathematics Supplementary Studies*, 1:167–212, 1978.

[28] M. S. Waterman. *Introduction to Computational Biology*. Chapman and Hall/CRC, 1995.

[29] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[30] S. Wuchty. Small worlds in RNA structures. *Nucleic. Acids. Res.*, 31(3):1108–1117, February 2003.

[31] T. Xia, Jr. J. SantaLucia, M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, and D.H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–35, 1999.